Democracy Distorted – Deepfakes as Political Weapons

Gjon Rakipi,[1] Yasaman Yousefi,[2, 3]
Calogero Caltagirone,[4] Angelo Tumminelli,[4]
Andrew McIntyre,[5] Aseniya Dimitrova[6]

[1] Albanian Institute for International Studies (AIIS), Tirana, Albania
gjonrakipi@aiis-albania.org
[2] DEXAI-Artificial Ethics, Rome, Italy
yasaman.yousefi@dexai.eu
[3] University of Bologna, CIRSFID ALMA AI, Faculty of Legal Studies, Bologna, Italy
y.yousefi@unibo.it
[4] LUMSA University, Department of Human Sciences, Rome, Italy
a.tumminelli@lumsa.it, c.caltagirone@lumsa.it
[5] University of Amsterdam, Institute for Logic, Language and Computation; Amsterdam, the Netherlands
a.mcintyre@uva.nl
[6] Brand Media Bulgaria, Sofia, Bulgaria
seniya.dimitrova@gmail.com

Affordable generative AI allows actors to produce and amplify deepfakes instantly, outpacing verification efforts. Drawing on Young's (2011) distinction between isolated harms and structural injustice, this chapter identifies synthetic media as a structural threat to democracy that collapses the evidentiary foundations of public reason. We examine how deepfakes weaponize information ecosystems, using European and U.S. case studies to demonstrate their specific deployment against women and minority candidates. Methodologically, we analyse recent disinformation incidents through the lenses of epistemic injustice and deliberative democracy. We argue that deepfakes signal a deeper vulnerability where truth becomes malleable and public trust erodes. The chapter concludes that safeguarding democratic life requires not only legal and technical fixes, but a normative reorientation toward truthfulness and accountability.

DOI
https://doi.org/10.18690/um.feri.2.2026.5

ISBN
978-961-299-109-8

Keywords:
disinformation, electoral integrity, epistemic injustice, gendered harms, democratic resilience

University of Maribor Press

1 Conceptualising Harm

Harm is an elastic idea. In its oldest sense, it names any blow to a person's well-being: a broken bone, a stolen wage, a silenced voice. Yet the digital century invites a broader lens. Today, a manipulated recording, such as the AI-generated audio targeting Michal Šimečka just days before Slovakia's 2023 vote (Meaker, 2023), can circulate in the morning, fracture public trust by noon, and tilt an election by evening. Such episodes remind us that harm is both material, and epistemic and political. Epistemic harm occurs when the channels through which we come to know the world are deliberately muddied. Deepfakes, coordinated rumour campaigns, and AI-generated "news" flood the evidentiary pool with noise, making it harder for individuals to sort fact from fabrication. Uncertainty is not a neutral by-product here; it is the intended wound, eroding a community's capacity to share reasons and reach common judgments. Political harm builds on this erosion. Democratic life depends on citizens who can verify, contest, and ultimately consent to the decisions made in their name. When falsehoods travel faster than rebuttals, accountability mechanisms falter. The result is not just misinformed voters but a weakening of the very norms that make collective self-government possible. By foregrounding these layered harms, the chapter can shift from cataloguing threats to explaining why they matter normatively, providing the conceptual framework we will use to analyse gendered disinformation (Section 5.4) and the erosion of democratic values (Section 5.5). Readers will see that the stakes extend beyond isolated victims to the cognitive and institutional scaffolding on which democratic societies rest.

2 Electoral Interference in Europe and Beyond

Elections are pivotal moments for democratic societies, where this single event can significantly alter power structures, policy directions, and political representation at local, national, and international levels. Both are the outcomes of elections highly consequential, but they also often trigger periods of intense political engagement and polarization among citizens, as competing socio-political messages come to the forefront of public discourse and debate. Additionally, elections are highly mediated events as political parties, and their supporters communicate their messages to the public via a wide range of media channels (Mazzoleni & Schulz 1999). This includes campaign materials (e.g., posters, adverts, leaflets), political activities (e.g., speeches, press conferences) and journalistic coverage (e.g., opinion pieces, interviews,

televised debates). This mediatization of elections has only intensified with the rise of social media platforms, wherein political content can be directly communicated to individual users in a highly personalized way through network connections, algorithmic recommendations, and targeted advertising (Marwick & Lewis 2017; Chun 2021).

The combination of highly consequential outcomes, a politically sensitive environment, and the pervasive mediation of political messaging means that elections are particularly attractive and vulnerable targets for political manipulation through coordinated disinformation campaigns. Given these factors, even the uncoordinated and/or unintentional spread of disinformation during election periods can have a significant impact.

With the arrival of modern generative AI systems and the widespread production and spread of synthetic media online, elections have become ever more dangerous times for democratic societies. Generative AI systems are now capable of producing high-quality synthetic audiovisual content (e.g., images, video, audio, text) that is near-indistinguishable from authentic content (Yazdani et al. 2025). Furthermore, the arrival of these systems means that the production of high-quality disinformation is less costly (Smith and Mansted 2020). Synthetic media depicting government officials, political figures and influential media personalities doing or saying anything could have a significant impact on the outcome of elections (Chesney & Citron, 2019; Diakopoulos & Johnson, 2019). For example, such content could be used to undermine the reputation of public figures, deceptively sway public opinion on specific issues, and/or threaten influential figures to manipulate their actions and political positions.

Since the emergence of deepfakes in 2017, there have already been numerous high-profile cases of synthetic media being used for electoral interference. For example, in the run up to the Slovak parliamentary elections in 2023, synthetic audio released online appeared to show politician Michal Šimečka, leader of the Progressive Slovakia party, discussing plans to rig the election in an attempt to undermine his credibility in the eyes of voters (Meaker, 2023). Meanwhile, the 2024 Pakistan general election saw several synthetic audiovisual recordings circulating online. These appeared to show prominent members of the Pakistan Tehreek-e-Insaf (PTI) party, including imprisoned leader Imran Khan, calling for a boycott of the election meant

to deceive PTI supporters into abstaining (Tiwari, 2024). In both cases, the synthetic content was identified as inauthentic by news media and the impact upon the election was seemingly minimal. Progressive Slovakia came second in the parliamentary elections, while in Pakistan PTI-backed candidates won more seats than any other single party. Ironically enough, Khan declared victory from jail using synthetic media. Though technically convincing, synthetic content that misrepresents high-profile political figures like Šimečka and Khan is unlikely to deceive a significant proportion of the public to have a considerable impact. This is because such content receives considerable attention and scrutiny to be easily detected and debunked. What is less widely discussed, but potentially more dangerous to electoral integrity, is the use of synthetic media in low-profile political settings; so-called "microfakes".

Where high-profile disinformation is likely to be debunked, synthetic content depicting figures and officials involved in smaller-scale politics may go undetected as such content is unlikely to be widely distributed and properly scrutinized (Ascott, 2020). Smaller-scale disinformation campaigns featuring local politicians or officials addressing local controversies (e.g., road quality, bypass development, cycle lanes) may appear technically convincing and interfere with local elections. Though there is currently little evidence of real-world microfakes, cases are unlikely to be reported by their very nature. As one clear example, during the 2022 mayoral election in Shreveport, Louisiana, the likeness of incumbent Democratic candidate Adrian Perkins was digitally recreated using AI as part of a hostile political advertisement criticising his policies (Swenson et al. 2024). Perkins ultimately lost the election and claims this deepfake advertisement played a crucial role. Though openly artificial and intended as humorous satire, this advertisement proves that such microfakes could be utilized at a local level. While the immediate impact of these microfakes may be minor, coordinated disinformation campaigns targeting numerous local elections could represent a granular and gradual threat to democracy that escalates to influence national and international politics.

Beyond disinformation campaigns aimed directly at undermining the credibility of candidates or influencing voter sentiments on specific issues, synthetic media can also be used to intimidate, threaten or otherwise harass political figures to influence their actions and statements, or to deter political participation altogether (Chesney & Citron, 2018). Notably, the production of deepfake pornographic content presents a significant reputational risk and thus the very threat of publication could

be used to deter candidates from standing in elections, as will be discussed in more detail below (Adjer et al., 2019; Rini & Cohen, 2022).

While the arrival of generative AI may be exacerbating risks for electoral interference, synthetic content emerged into an information environment that was already fertile ground for rampant disinformation and post-truth politics. Throughout the 2010s and into the 2020s, there has been a noted decline in traditional news media as people have grown more dependent on social media platforms as the primary source of political information. Unlike traditional journalism which relies on editorial standards and fact-checking, social media platforms operate and disseminate content according to an attention economy wherein there is such an overabundance of content that the flow of information hinges upon what will attract people's attention immediately (Lewis & Marwick, 2017). Such a system prioritizes emotionally charged or sensational content rather than complex, nuanced information. More so than traditional media. As such, these networks allow for disinformation and false narratives to circulate widely among platform users before traditional journalists and fact-checkers can publish evidence-based rebuttals or corrections. Within this attention economy, sensational political synthetic media may spread online too rapidly or go entirely unnoticed, potentially influencing users that have little media literacy skills or that are less engaged with broader political discourse and debates. These networks are also extremely vulnerable to attention-hacking techniques that seek to manipulate those content filtering and recommendation algorithms that dictate what information users see and interact with. For example, throughout the 2010s, far-right extremists frequently coordinated large groups of users to flood Twitter with specific hashtags (e.g., #gamergate) to artificially make this topic trend and reach users who might not otherwise encounter their propaganda. In other instances, these extremists have piggybacked on already trending hashtags (e.g., #blacklivesmatter) to hijack its popularity and strategically amplify the reach of their own political messages.

Designed to capitalise on this attention economy, algorithmic recommendation systems preferentially show users content that provokes engagement. In doing so, these systems reinforce pre-existing biases and deepen divisions along ideological lines. Building on this algorithmic polarization, users of online platforms are increasingly connected based on the principle of homophily i.e., the assumption that similarity breeds connection (Chun, 2024). Algorithmic recommendation systems

cluster individual users into neighbourhoods based on similarity (e.g., race, gender, sexuality, political affiliation). This clustering encourages political echo chambers to form wherein there is little exposure to conflicting information and people are encouraged to accept information that confirms their existing beliefs, regardless of its accuracy. Within such neighbourhoods, political messaging and disinformation can spread freely and with greater impact via strong interpersonal ties among members. Synthetic content promoting false political narratives can, therefore, be more readily accessed, accepted and shared. Once embedded, these false narratives are difficult to combat, shaping voter perceptions and undermining trust in the legitimacy of democratic societies.

More generally, the proliferation of synthetic media that is near-indistinguishable from authentic content means that people are more sceptical of all information they receive online, and they are less likely to trust traditional information sources and authorities (Vaccari & Chadwick, 2020). The epistemic impact of synthetic media on our information environment more broadly is discussed in the next section.

3 Epistemic Erosion and the Misinformation Ecosystem

Beyond headline elections, deepfakes exacerbate the chronic "liar's dividend": the mere possibility that any footage might be fabricated empowers bad actors to dismiss authentic evidence and fuels public cynicism. A 2024 European Parliamentary briefing warns that synthetic media risks a downward spiral in which voters "no longer believe what they see or hear," undermining media pluralism and parliamentary scrutiny (Michael & Hocquard, 2023). UNESCO's report (2023) on freedom of expression during elections similarly notes that cheap-fakes and deepfakes erode basic informational rights by diffusing responsibility among anonymous creators, automated recommender systems, and inattentive platforms. Experimental work published in Digital Journalism finds that high-quality deepfakes reduce viewers' trust in both the target and the outlet that hosts the correction, even when the fabrication is revealed within seconds (Patel, 2025). The study referenced, published in the journal Digital Journalism, is part of a growing body of research examining the impact of deepfakes on public trust. Deepfakes are AI-generated manipulated videos capable of producing extremely realistic footage, often difficult to distinguish from authentic content. The researchers conducted controlled experiments in which participants were shown short, high-quality deepfake videos,

followed by an immediate correction or debunk published by a news outlet. The interval between viewing the deepfake and being informed of its falsity was only a few seconds, an intentionally "ideal" scenario in which both the victim and the news organization respond as quickly and transparently as possible. The cumulative outcome is an epistemic environment where strategic actors can manufacture plausible doubt faster than institutions can generate consensus, eroding the public's capacity for informed deliberation.

3.1 Infodemic and Epistemic Erosion: The Role of Deepfakes

An infodemic is a phenomenon in which an excessive amount of unverified or contradictory information makes it difficult for recipients to ground themselves in reality (World Health Organization, 2020; Cinelli et al., 2020). The category of "infodemic" has gained importance, especially during the COVID-19 pandemic, but it represents a broader and ongoing issue that is linked to the digital age in which news, true, false, or distorted, spreads at unprecedented speeds.

This is the context in which a subset of generative AI known as deepfakes emerges. Deepfakes are able to bolster the infodemic, making it increasingly difficult to distinguish between what is authentic and what is manipulated. Their impact is both informative and epistemic in that they undermine our ability to trust traditional sources and media, reconfiguring the very modalities of knowledge and perception of the world.

This epistemic erosion weakens the pact of trust on which shared knowledge is based. In fact, when even digital content can be manipulated in a dystopian way, our perception of reality itself becomes fragile and fuels an informational relativism that opens the doors to a dangerous revisionism and systemic distrust.

Without critical tools and adequate regulatory frameworks, we risk having a society in which the truth is not only manipulable but also completely delegitimized. To counter this drift, it is necessary to invest in media literacy and accountability.

AI certainly represents one of the most insidious challenges for public information in the 21st century: it is a non-neutral tool that, if used maliciously, can become a powerful vehicle for disinformation and epistemic dystopia. In fact, in public contexts, such as politics, journalism, or social debate, deepfakes undermine the

reliability of content and contribute to eroding truth as the foundation of collective discourse (Weikmann & Lecheler, 2024). This determines the phenomenon that has been appropriately defined as "epistemic pollution" with which information is distorted, manipulated or presented in a misleading way, compromising our ability to know and understand the world (Levy, 2021). In a dystopian context, the use of artificial intelligence can amplify this phenomenon, generating intentionally false but credible content. Algorithms trained on partial or manipulated data can reinforce pre-existing biases, creating information bubbles and cognitive polarization (Praiser, 2011; O"Neil, 2016). This phenomenon fuels a dangerous form of information nihilism (Labarre, 2025), in which every truth is suspect, every piece of evidence is revocable, and every opinion becomes equally valid. In such a climate, truth loses its value and illusion takes over. The consequences are profound: social polarization, civic disillusionment, and the delegitimization of democracy. Furthermore, and very relevant to this reflection, AI can be used by authoritarian regimes or interest groups to rewrite historical and cultural narratives (Hameleers et al., 2024). In the absence of transparency and control, reliable sources lose relevance, and access to knowledge is filtered by opaque interests. Information democracy turns into an algorithmic oligarchy that must be countered through critical awareness and the ethical governance of AI.

Addressing the impact of deepfakes requires rethinking verification standards, promoting digital literacy, and holding content creators and platforms accountable. Only through these efforts can truth be defended in an increasingly vulnerable public sphere.

3.2 The ethical dimensions of deepfakes

Deepfakes blur the line between authentic and fabricated evidence, threatening individual autonomy and public trust. This has serious implications for fields like journalism and law enforcement, where visual evidence plays a critical role. Fabricated content in these areas can have far-reaching consequences, including the corruption of the historical record, the miscarriage of justice, and the undermining of public trust in essential institutions. The issue of consent is also paramount when it comes to deepfakes. Using someone's likeness without their agreement, particularly for harmful purposes, violates personal rights and dignity. The potential use of deepfakes in international relations adds another layer of complexity to the

ethical debate. They could be used to create false evidence, to mislead the public or international community, and potentially to provoke conflicts or exacerbate

4 Gendered and Minority Harms

As discussed in earlier sections, the advent and diffusion of synthetic media technologies, particularly deepfakes, pose significant challenges to democratic life. However, it is essential to recognize that these harms are not borne equally. An emerging body of evidence demonstrates that the impacts of deepfakes are disproportionately experienced by women and minority groups, both in their private lives and in the public sphere. This section examines how deepfakes operate as technological amplifiers of entrenched social inequalities, drawing on empirical research, legal scholarship, and documented case studies to articulate their normative and political consequences.

A pivotal moment in this discourse came with the 2019 audit conducted by the cybersecurity firm Deeptrace. Their findings revealed that 96 percent of the 14,678 deepfake videos indexed at that time were non-consensual pornographic content targeting women (Adjer et al., 2019). Subsequent studies have since corroborated this troubling trend. For instance, a 2024 survey spanning ten countries found that 2.2 percent of respondents reported being targeted by synthetic intimate imagery without their consent, with women and gender minorities disproportionately represented among the victims (Umbach et al., 2024). These figures illustrate a broader phenomenon: the weaponization of deepfake technology to perpetuate gender-based violence and harassment.

While the development of generative AI was initially confined to research circles, this changed in 2017 when a Reddit user under the pseudonym "Deepfakes" began distributing manipulated pornographic videos using free, open-source machine learning tools. This marked a turning point in the accessibility and misuse of synthetic media, setting a precedent for widespread abuse.

Academic literature has repeatedly emphasized the gendered nature of deepfake harms. Chesney and Citron have argued that non-consensual deepfake pornography, as one of the earliest and most prevalent applications of the technology, systematically targets women and introduces novel forms of gender-based abuse. With minimal technical expertise, perpetrators can now fabricate highly realistic

sexual content using another person's likeness, thereby enabling a continuum of exploitative practices that includes sextortion, reputational sabotage, blackmail, and intimate partner violence (Chesney & Citron, 2018). Yet the scope of exploitation is not limited to sexualized media. Deepfakes have also been deployed in cases of identity fraud, financial scams, and emotional coercion, including fabricated kidnapping videos or synthetic recordings designed to manipulate or intimidate. These forms of abuse are not merely technological anomalies; they reflect deeper structural patterns in which individuals are rendered tools for others' gain, often at great personal and societal cost.

Laffier and Rehman have further highlighted the psychological and reputational consequences of these abuses, noting that victims frequently suffer job loss, social exclusion, and severe mental health outcomes (Laffier & Rehman, 2023). The weaponization of deepfakes against women and minority communities thus functions as a form of personal attack and as a mechanism for reinforcing existing social hierarchies and exclusions.

In political contexts, these harms have a particularly corrosive effect on democratic participation. Deepfakes increasingly operate as tools of deterrence, strategically targeting underrepresented groups to dissuade them from civic engagement. They undermine the democratic ideal of equal participation by selectively amplifying social vulnerabilities and exploiting pre-existing prejudices. Female politicians, already the subject of disproportionate online abuse, now contend with the added threat of AI-generated disinformation. Such campaigns are capable of producing fabricated pornographic material, falsified news articles, and synthetic audiovisual recordings, all designed to erode credibility and sow distrust.

One of the most troubling aspects of gendered disinformation is its adaptability. Algorithmic systems can customize fabricated content to match the biases of particular audiences (Goldstein et al., 2023). In conservative-leaning electorates, such content may depict women in line with regressive gender stereotypes, questioning their emotional stability or capacity for leadership. In more progressive regions, false narratives may be engineered to simulate scandal or ethical misconduct. Regardless of context, the end goal remains the same: to undermine a woman's professional and political legitimacy.

The deployment of deepfakes in electoral politics is increasingly well-documented. In France, ahead of the 2024 EU elections, deepfake videos circulated online purporting to show young women identified as nieces of Marine Le Pen endorsing far-right ideologies. These videos, though fabricated, gained significant traction and sparked renewed debate over the inadequacy of content moderation in responding to political disinformation (Hartmann, 2024). In Germany, during the 2021 federal election, Annalena Baerbock, the Green Party's candidate for Chancellor, was the target of AI-generated narratives laced with gendered tropes and intimidation tactics. These efforts compromised her individual campaign, and sent a chilling message to women contemplating political careers (Kovalčíková & Weiser, 2021). In Italy, female politicians across the political spectrum, including Prime Minister Giorgia Meloni and opposition leader Elly Schlein, have been targeted with deepfake pornography and sexually explicit images, forming part of a broader strategy of delegitimization through misogynistic content (Chopra et al., 2025; Giuffrida, 2025).

These attacks are part of a broader strategy of participatory deterrence. By inflating the reputational and personal costs of public life, deepfakes serve to exclude marginalized groups from democratic institutions. The concept of epistemic injustice, as theorized by Miranda Fricker, proves useful here, specifically her notion of 'testimonial injustice,' which describes how prejudice leads audiences to assign a 'credibility deficit' to a speaker, wrongly stripping them of their status as a reliable knower. (Fricker, 2007). It captures the systematic devaluation of certain groups as credible knowers and participants in public discourse. Deepfakes exacerbate such injustice by selectively targeting those who already face structural disadvantages, thereby intensifying their marginalization. The result is an informational environment in which appearances override evidence, and democratic deliberation gives way to aesthetic manipulation, echoing concerns about an emerging "post-truth geopolitics" (Chesney & Citron, 2019).

A further challenge lies in the responses, or lack thereof, by digital platforms. Social media companies and content-sharing platforms often treat pornographic deepfakes as privacy issues rather than as democratic threats. Consequently, moderation and takedown mechanisms tend to lag behind the speed at which such content spreads, allowing politically motivated synthetic media to reach wide audiences before fact-checkers can intervene (Chesney & Citron, 2018). This regulatory inertia enables malicious actors to exploit algorithmic amplification and virality, often with impunity.

The harm is amplified by the architecture of digital platforms themselves. Deepfakes can be created with basic tools, uploaded in seconds, and rapidly disseminated across networks at little to no cost. Victims and public institutions frequently struggle to keep pace. Even after content is debunked, its reputational damage often persists, illustrating the profound temporal and institutional asymmetries embedded in the current media ecosystem.

Compounding this situation is a failure of governance. Carpenter notes cheap-fakes and deepfakes fracture the informational commons by diffusing accountability across anonymous creators, automated content delivery systems, and disengaged platform policies. The result is an epistemic landscape where both truth and trust are undermined, and where the mere possibility of fabrication, the so-called "liar's dividend", is sufficient to discredit even authentic evidence (Carpenter, 2024).

In sum, the gendered and minority harms of deepfakes are not isolated incidents but structural phenomena that exploit existing inequalities, distort democratic processes, and degrade informational integrity. Addressing these harms demands, at a superficial level, technical fixes and, more profoundly, a normative reorientation that centres justice, accountability, and inclusive participation in the governance of emerging technologies.

5 Normative Implications for Democratic Values

Liberal democracy relies on citizens being able to verify what leaders say and do. When a convincing AI-generated video or audio circulates, that shared evidentiary ground can disappear. Deliberative theorists such as John Rawls describe this ground as the basis of public reason, the arena where disagreements are settled with facts that everyone can inspect. Deepfakes undermine that arena in two reinforcing ways. First, they insert persuasive falsehoods faster than journalists and fact checkers can react. Second, the very existence of generative forgeries lets dishonest actors deny authentic evidence. This forementioned liar's dividend means that someone caught in wrongdoing can claim the incriminating video is merely synthetic (Chesney & Citron, 2018). Both dynamics erode transparency because they make visual or auditory proof negotiable rather than authoritative.

The European Union's Artificial Intelligence Act (Article 50) will require clear labelling of synthetic audiovisual content to restore minimum transparency, but enforcement will not begin until the regulation's phased entry into force in 2025 (*European Union Artificial Intelligence Act: A Guide*, 2025). Until then, Europeans inhabit what philosopher Regina Rini describes as an epistemic fog where seeing is no longer believing.

Democracy promises that every citizen's contribution deserves comparable credibility. Deepfakes threaten this promise by amplifying pre-existing asymmetries of capacity and access. Producing convincing synthetic media still demands specialized skills, substantial computing power, or paid software, whereas evaluating authenticity usually requires time, digital literacy, and sometimes proprietary forensic tools. Well-resourced actors, for example, large campaigns, state broadcasters, or private influence firms, therefore, enjoy a comparative advantage in shaping narratives, while ordinary citizens must consume content in real time without equivalent verification resources. One 2019 article notes that deepfake operations concentrate communicative power in the hands of those with technical sophistication, and such a concentration is able to skew public deliberation toward elites with asymmetric informational control (Kietzmann et al., 2020).

From a deliberative perspective, the problem is not simply unequal speech volume, but unequal credibility allocation. Citizens lacking digital-forensic literacy are more likely to accept forged media as real or to dismiss genuine media as fake, creating what epistemologists describe as credibility deflation, a systemic reluctance to trust anyone who lacks signals of technological authority. Rural populations, older voters, and linguistic minorities often face additional barriers to reliable verification services, perpetuating a civic hierarchy in which those with access to advanced tools can define what counts as knowledge. Equality suffers even without targeted harassment because the communicative space tilts toward actors who can purchase sophisticated deception or rapid authentication.

Transparency failures and credibility gaps combine to weaken accountability, the process that turns democratic judgment into real consequences. Deepfakes enable false scandals to destroy reputations overnight, and let genuine misconduct be waved away as "fake" procedures meant to encourage calm reflection can be hijacked by synthetic evidence that spreads suspicion when replies are legally muted.

Jürgen Habermas stresses that democratic legitimacy rests on communicative rationality, a norm requiring actors to justify their positions with reasons subject to public testing. Deepfakes loosen the bond between action and proof, enabling officials to evade substantive answers by questioning the medium itself. The public sphere risks sliding toward post-truth politics, a climate in which empirical validation yields to partisan loyalty.

Transparency, equality, and accountability form an interlocking architecture. When transparency falters, resource-rich actors exploit the uncertainty, which deepens inequality in communicative power. That inequality then makes it easier for influential players to deploy or dismiss synthetic media, further weakening accountability. Scholars of systemic deliberative democracy emphasise that legitimacy arises from the composite health of these channels rather than isolated exchanges. Deepfakes compromise the channels simultaneously, creating a spiral in which each weakened pillar accelerates the decay of the others.

Europe's nascent responses acknowledge this systemic threat but remain partial. Labelling mandates in the AI Act aim to shore up transparency, while proposed platform-researcher partnerships under the European Democracy Action Plan seek to democratise verification capacity, thereby easing equality gaps. Finland's National Media Education Policy (2019) emphasizes systematic media education, quality, and lifelong learning, linking it to societal resilience in the face of disinformation threats (Finland, 2024).

Yet norms must evolve alongside laws. Deliberative legitimacy depends on civic cultures that prize truthful presentation, reciprocal respect, and willingness to be answerable. Technical interventions can scaffold those virtues, but they cannot substitute for them.

Deepfakes expose a vulnerability at the core of democratic architecture, where authenticity functions as a prerequisite for collective self-government. By destabilising what counts as evidence, concentrating communicative power, and enabling strategic denial, synthetic media corrodes the normative pillars that make democracy possible. Regulatory measures may restore partial transparency, and educational programs may narrow literacy gaps, yet democracy ultimately survives on public commitments to truth, equal regard, and responsibility. Reaffirming these

commitments in an era of perfect forgeries is not peripheral to technology policy; it is central to democratic renewal.

6 Policy and Educational Responses

Generative-AI systems already create text, images, video, and audio that are almost indistinguishable from authentic material, and the European Commission's Generative AI Outlook warns that such synthetic content could erode public trust during elections and crises if safeguards, including both provenance tracking to verify origin and forensic detection to identify manipulation, do not keep pace (Navajas Cawood et al., 2025). Legislators and regulators are therefore moving from aspirational principles to binding rules that criminalise harmful deepfakes, require visible labelling or watermarking, guarantee rapid takedown mechanisms, limit synthetic political advertising, place detection duties on intermediaries, and oblige model developers to publish transparency reports on training data and risk controls.

Inside the European Union, Article 35 of the Digital Services Act obliges very large online platforms to assess and mitigate systemic risks from manipulated media, label AI-generated content, and give independent researchers secure audit access, with penalties of up to six percent of worldwide turnover for non-compliance. A strengthened Code of Conduct on Disinformation, now formally linked to the Act, extends similar transparency and risk-mitigation duties to search engines and social networks of all sizes and tightens rules on political advertising that uses generative. Forthcoming obligations in the AI Act will reinforce that framework by requiring anyone who publishes synthetic images, audio, or video depicting real people to add notices readable by humans and machines.

Several member states have already gone further. Spain empowered its AI authority to levy fines of up to €35 million, or seven percent of global turnover, on platforms that fail to label synthetic content clearly.[1] France amended its Penal Code to prohibit distributing deepfakes that use a person's likeness or voice without consent unless the artificial origin is disclosed, imposing tougher penalties for sexual material or large-scale online dissemination (Coslin et al., 2024). Commentators note that the new article gives prosecutors a versatile weapon against disinformation campaigns and celebrity impersonations. Germany's Bundesrat circulated a draft Digital

[1] See https://digital-strategy.ec.europa.eu/en/library/code-conduct-disinformation

Forgery Act that would criminalise synthetic impersonation and introduce higher penalties when victims suffer reputational or economic harm (*Germany: Bundesrat Publishes Draft Law on Deepfakes | News*, 2024). Denmark proposes a copyright-style right in personal biometric features, so reproducing a face or voice in artificial media would require permission or risk infringement liability (Bryant, 2025).

The Italian Constitution safeguards personality rights, including the right to control one's image. Additionally, the Italian Civil Code in its Article 10.5 prohibits the unauthorized use of an individual's likeness, and personal data legislation also protects this. These laws, along with the Italian Copyright Law, enable individuals to seek compensation if their image is used without their consent, especially if it harms their honour or reputation. This clause can arguably be extended to the use of deepfakes. A notable case that exemplifies the enforcement of these laws involved Prime Minister Giorgia Meloni in a lawsuit over pornographic synthetic videos viewed millions of times (Gozzi, 2024). The United Kingdom's Online Safety Act criminalizes both sharing and creating non-consensual intimate deepfakes, with unlimited fines and possible prison sentences (BCC, 2023).

In the United States, the federal landscape remains fragmented, but Congress has introduced the TAKE IT DOWN Act to criminalise non-consensual intimate deepfakes nationwide and compel platforms to provide expedited removal tools (Sen. Cruz, 2025). States continue to fill gaps. Alabama's Child Protection Act treats AI-generated sexual imagery involving minors as virtually indistinguishable from real abuse material (*Alabama HB168*, 2024). California extended its post-mortem right of publicity so that distributing a digital replica of a deceased person without consent triggers civil liability and statutory fines (Wolff & Safran, 2024). Alabama also adopted a Materially Deceptive Election Media statute that outlaws AI-generated content intended to mislead voters (Guidry & Amin, 2024). Arizona clarified that its intimate-image law covers synthetic as well as genuine photographs (Ventura, 2024). Digital Identity Theft Act obliges platforms to host a simple tool for victims, especially minors, to remove explicit deepfakes and criminalizes their non-consensual creation or distribution (*Senator Wahab's Stop the Online Predators Act and Digital Identity Theft Act Signed into Law*, 2024).

Multilateral coordination began to crystallise with the Hiroshima AI Process, whose guiding principles urge developers to publish capability cards, specify disallowed uses, protect intellectual property, and invest in user education so citizens can

recognise synthetic media (Japan Gov, 2024). Further to that, the Bletchley Declaration from November 2023 commits signatories to cooperate internationally on the safe development, deployment, and governance of powerful "frontier" AI systems (Department for Science, Innovation and Technology, 2023). Yet implementation is uneven: a European Digital Media Observatory evaluation for the first half of 2024 found that very large platforms met many Code-of-Practice labelling and removal commitments, but smaller services showed limited engagement and inconsistent reporting, underscoring the need for enforcement and capacity-building (Botan & Meyer, 2025).

Because legislation alone cannot keep pace with rapidly improving models, policymakers emphasise technical safeguards and education. A European Parliament briefing on children and deepfakes calls for age-appropriate curricula that teach pupils, parents, and teachers to evaluate digital sources, recognise emotional manipulation, and use verification tools (Negreiro, 2025). The OECD, in cooperation with the European Commission, is drafting an AI-literacy framework that will guide the next Programme for International Student Assessment cycle and provide lesson plans on generative AI (Schleicher, 2025). At the civic level, the EU-funded EUvsDisinfo platform offers an open database of disinformation narratives, interactive games, and instructional videos that help users practise source checking and critical reading (*About - EUvsDisinfo*, 2025).

Research agencies and private companies invest heavily in detection. In the United States, DARPA funds the Semantic Forensics and Media Forensics programmes, which develop algorithms to spot compression artefacts, lighting inconsistencies, and biometric mismatches that indicate tampering (*SemaFor: Semantic Forensics | DARPA*, 2025). Midjourney, a major generative-image service, voluntarily blocks prompts that attempt to create pictures of prominent political figures during election periods, reducing the risk of deceptive visuals entering public debate (O"Brien, 2024). The Massachusetts Institute of Technology's Detect DeepFakes project provides an online training tool where users test their ability to identify manipulated material, and researchers measure how such exercises improve resistance to misinformation.[2] Finland complements these efforts by integrating media-literacy instruction from primary school onward and pairing classroom exercises with

[2] See link: https://detectfakes.media.mit.edu/

public-service broadcasts that explain how manipulated content spreads and how to debunk it (Finland, 2024).

Together, these initiatives raise the cost of deception while preserving the legitimate benefits of generative AI. The European Union's layered strategy, combining horizontal rules like the Digital Services Act with national adaptations and ongoing sector-specific reforms, illustrates how a comprehensive framework can emerge without stifling innovation. In the United States, federal and state measures show that even a patchwork can converge on core principles of consent, transparency and rapid redress. Multilateral dialogues, voluntary industry standards, open-source detection tools and grassroots media-literacy campaigns complete this defence, giving citizens the knowledge and technical support they need to judge what they see and hear before sharing it.

Despite the differences in national approaches, all reviewed examples share the common goal of curbing deepfake abuses and safeguarding the dignity and personal data of citizens. The successful implementation of the European framework (mainly EU AI Act) as a first comprehensive attempt, followed by national adaptations, is expected to lead to more strict enforcement and oversight, and on the flexibility to respond to rapid technological developments. In this context, coordinated international dialogue and the exchange of best practices among Member States are crucial to achieving a balanced and effective regulatory approach that combines innovation with the protection of fundamental human rights.

Beyond formal legislation, industry-led guidelines, technological safeguards, and public awareness campaigns play a vital role in mitigating deepfake risks and promoting responsible AI use. These initiatives, ranging from open-source detection tools to media literacy programs, complement regulatory frameworks by fostering grassroots resilience and rapid adaptation to emerging threats.

Ultimately, a holistic strategy that combines binding rules with voluntary standards and civil-society engagement offers the best path toward an ecosystem where innovation thrives under robust ethical guardrails.

7 Concluding Remarks

Deepfakes pose a structural threat to democratic life by destabilizing the evidentiary foundations of public reason, accountability, and trust. They accelerate the spread of falsehoods while enabling the denial of authentic evidence, creating an epistemic environment in which citizens struggle to distinguish truth from fabrication. These dynamics disproportionately affect women, gender minorities, and other marginalized groups, amplifying social inequalities and deterring full participation in civic and political life. By concentrating communicative power in the hands of technologically sophisticated actors, deepfakes exacerbate inequalities in credibility and reinforce structural hierarchies, undermining the democratic ideal of equal participation.

Addressing these challenges requires a multi-layered approach combining legislation, platform regulation, technological safeguards, and education. Policies such as mandatory labelling, rapid takedowns, and penalties for harmful content, alongside media literacy programs and detection tools, help citizens navigate an increasingly complex information ecosystem. We acknowledge, however, that this analysis is limited by the nascent stage of these regulatory frameworks, whose long-term efficacy in curbing algorithmic disinformation remains to be empirically tested. Yet legal and technical measures alone are insufficient: the resilience of democracy ultimately depends on nurturing civic norms of truthfulness, accountability, and inclusive participation. To this end, future research should prioritize empirical studies that measure the long-term impact of specific media literacy interventions on citizen resilience across diverse political environments.

End notes

Gjon Rakipi conceptualized the chapter, contributed to the abstract, and wrote the sections "Conceptualising Harm" and "Normative Implications for Democratic Values". Additionally, he carried out the preliminary full-chapter edit and assisted in referencing. Andrew McIntyre is the author of "Electoral interference in Europe and Beyond". Calogero Caltagirone and Angelo Tumminelli co-authored "Epistemic Erosion and the Misinformation Ecosystem." Yasaman Yousefi refined the abstract and wrote "Gendered and Minority Harms," drafted the conclusion, performed the final edit, and assisted with the referencing. Asenia Dimitrova authored "Policy and Educational Responses." All authors were individually responsible for the literature review and writing of their respective sections.

References

EUvsDisinfo. (2025, October 27). *EUvsDisinfo*. https://euvsdisinfo.eu/about/

Adjer, H., Giorgio, P., Francesco, C., & Laurence, C. (2019). *The state of deepfakes: Landscape, threats, and impacts*.

Alabama HB168. (2024, May 2). *TrackBill*. https://trackbill.com/bill/alabama-house-bill-168-crimes-offenses-raises-max-age-for-offenses-involving-obscene-materials-with-depictions-of-children-authorizes-punitive-damages-for-victims-of-those-offenses-and-directs-board-of-ed-to-require-policies-related-to-those-offenses/2517763/

Ascott, T. (2020). Microfake: How small-scale deepfakes can undermine society. *Journal of Digital Media and Policy, 11*(2), 215–222.

BBC. (2023, December 9). *AI: EU agrees landmark deal on regulation of artificial intelligence*.https://www.bbc.com/news/world-europe-67668469

Botan, M., & Meyer, T. (2025). *Implementing the EU Code of Practice on Disinformation: An evaluation of VLOPSE compliance and effectiveness (Jan–Jun 2024)*. EDMO. https://edmo.eu/publications/implementing-the-eu-code-of-practice-on-disinformation-an-evaluation-of-vlopse-compliance-and-effectiveness-jan-jun-2024/

Bryant, M. (2025, June 27). Denmark to tackle deepfakes by giving people copyright to their own features. *The Guardian*. https://www.theguardian.com/technology/2025/jun/27/deepfakes-denmark-copyright-law-artificial-intelligence

Carpenter, P. (2024). *FAIK: A practical guide to living in a world of deepfakes, disinformation, and AI-generated deceptions*. John Wiley & Sons.

Chesney, R., & Citron, D. (2019, February). *Deepfakes and the new disinformation war: The coming age of post-truth geopolitics*. Foreign Affairs. https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war

Chesney, R., & Citron, D. K. (2018). Deep fakes: A looming challenge for privacy, democracy, and national security. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3213954

Chopra, A., Masroor, G., & Blundy, R. (2025, January 6). "Form of violence": Across globe, deepfake porn targets women politicians. *France 24*. https://www.france24.com/en/live-news/20250106-form-of-violence-across-globe-deepfake-porn-targets-women-politicians

Chun, W. H. K. (2024). *Discriminating data: Correlation, neighborhoods, and the new politics of recognition*. MIT Press.

Cinelli, M., et al. (2020). The COVID-19 social media infodemic. *Scientific Reports, 10*, 16598.

Coslin, C., Gateau, C., & de Kouchkovsky, A. (2024, July 15). France prohibits non-consensual deep fakes. *Hogan Lovells*. https://www.hoganlovells.com/en/publications/france-prohibits-non-consensual-deep-fakes

Department for Science, Innovation and Technology. (2023, November 1). *The Bletchley Declaration by countries attending the AI Safety Summit, 1–2 November 2023*. GOV.UK. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

Diakopoulos, N., & Johnson, D. (2019). Anticipating and addressing the ethical implications of deepfakes in the context of elections (SSRN Scholarly Paper No. 3474183). *Social Science Research Network*.https://doi.org/10.2139/ssrn.3474183

European Union Artificial Intelligence Act: A guide. (2025, April 7). *Bird & Bird LLP*. https://www.twobirds.com/-/media/new-website-content/pdfs/capabilities/artificial-intelligence/european-union-artificial-intelligence-act-guide.pdf

Finland Ministry of Education and Culture. (2024, March 12). *Media literacy and education in Finland*. Finland Toolbox. https://toolbox.finland.fi/life-society/media-literacy-and-education-in-finland/

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198237907.001.0001

Germany: Bundesrat publishes draft law on deepfakes. (2024, July 9). *DataGuidance*.https://www.dataguidance.com/news/germany-bundesrat-publishes-draft-law-deepfakes

Giuffrida, A. (2025, August 28). Outrage in Italy over porn site with doctored images of prominent women. *The Guardian*. https://www.theguardian.com/world/2025/aug/28/outrage-in-italy-over-pornographic-website-with-doctored-images-of-prominent-women-giorgia-meloni

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). *Generative language models and automated influence operations: Emerging threats and potential mitigations*. arXiv. https://doi.org/10.48550/arXiv.2301.04246

Gozzi, L. (2024, March 20). Giorgia Meloni: Italian PM seeks damages over deepfake porn videos. *BBC News*.https://www.bbc.com/news/world-europe-68615474

Guidry, T., & Amin, T. (2024, May 24). Alabama takes a stand against AI in political campaigns. *The National Law Review*. https://natlawreview.com/article/new-ai-law-alert-alabama-next-state-take-stand-against-ai-generated-deceptive-media

Hameleers, M., van der Meer, T. G. L. A., & Dobber, T. (2024). They would never say anything like this! Reasons to doubt political deepfakes. *European Journal of Communication, 39*. https://doi.org/10.1177/02673231231184703

Hartmann, T. (2024, April 16). Viral deepfake videos of Le Pen family remind that content moderation is still not up to par ahead of EU elections. *Euractiv*. https://www.euractiv.com/news/viral-deepfake-videos-of-le-pen-family-reminder-that-content-moderation-is-still-not-up-to-par-ahead-of-eu-elections/

Japan Government. (2024, February 9). *The Hiroshima AI Process: Leading the global challenge to shape inclusive governance for generative AI*. https://www.japan.go.jp/kizuna/2024/02/hiroshima_ai_process.html

Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat? *Business Horizons, 63*(2), 135–146. https://doi.org/10.1016/j.bushor.2019.11.006

Kovalčíková, N., & Weiser, M. (2021, August 30). Targeting Baerbock: Gendered disinformation in Germany's 2021 federal election. *Alliance for Securing Democracy*. https://securingdemocracy.gmfus.org/targeting-baerbock-gendered-disinformation-in-germanys-2021-federal-election/

Labarre, J. (2025). Epistemic vulnerability: Theory and measurement at the system level. *Political Communication, 42*(1), 6–26. https://doi.org/10.1080/10584609.2024.2363545

Laffier, J., & Rehman, A. (2023). Deepfakes and harm to women. *Journal of Digital Life and Learning, 3*(1), 1–21. https://doi.org/10.51357/jdll.v3i1.218

Levy, N. (2021). Epistemic pollution. In N. Levy (Ed.), *Bad beliefs: Why they happen to good people* (Chap. 5). Oxford University Press. https://doi.org/10.1093/oso/9780192895325.003.0005

Lewis, B., & Marwick, A. E. (2017). *Media manipulation and disinformation online*. Data & Society Research Institute. https://datasociety.net/library/media-manipulation-and-disinfo-online/

Meaker, M. (2023, October 3). Slovakia's election deepfakes show AI is a danger to democracy. *WIRED*.https://www.wired.com/story/slovakias-election-deepfakes-show-ai-is-a-danger-to-democracy/

Michael, A., & Hocquard, C. (2023). Artificial intelligence, democracy and elections. *Artificial Intelligence*.

Navajas Cawood, E., Abendroth-Dias, K., Arias Cabarcos, P., Kotsev, A., Bacco, M., Bassani, E., Van Bavel, R., … Sellitto, A. (2025). *Generative AI outlook report: Exploring the intersection of technology, society, and policy*.Publications Office of the EU. https://data.europa.eu/doi/10.2760/1109679

Negreiro, M. (2025). *Children and deepfakes*.

O"Brien, M. (2024, March 13). AI image-generator Midjourney blocks images of Biden and Trump as election looms. *PBS News*. https://www.pbs.org/newshour/politics/ai-image-generator-midjourney-blocks-images-of-biden-and-trump-as-election-looms

O"Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin Press.

Patel, A. (2025). *Freedom of expression, artificial intelligence and elections*. UNESCO Digital Library. https://unesdoc.unesco.org/ark:/48223/pf0000393473

Rini, R., & Cohen, L. (2022). Deepfakes, deep harms. *Journal of Ethics and Social Philosophy, 22*(2). https://doi.org/10.26556/jesp.v22i2.1628

Schleicher, A. (2025, April 29). New AI literacy framework to equip youth in an age of AI. *OECD Education and Skills Today*. https://oecdedutoday.com/new-ai-literacy-framework-to-equip-youth-in-an-age-of-ai/

SemaFor: Semantic Forensics | DARPA. (2025, October 27). https://www.darpa.mil/research/programs/semantic-forensics

Sen. Cruz, T. (2025, May 19). *S.146 – TAKE IT DOWN Act (2025–2026)* [Legislation]. https://www.congress.gov/bill/119th-congress/senate-bill/146

Senator Wahab's Stop the Online Predators Act and Digital Identity Theft Act signed into law. (2024, September 19). *California Senate District 10*. https://sd10.senate.ca.gov/news/senator-wahabs-stop-online-predators-act-and-digital-identity-theft-act-signed-law

Swenson, A., Merica, D., & Burke, G. (2024, June 17). AI experimentation is high risk, high reward for low-profile political campaigns. *Associated Press*. https://apnews.com/article/election-2024-ai-deepfakes-political-campaigns-056c2200836e755826fbc9698bcfed60

Tiwari, S. (2024, February 8). Deepfakes become political weapon in Pakistan elections. *India Today*.https://www.indiatoday.in/world/story/deepfakes-become-political-weapon-in-pakistan-elections-2499399-2024-02-08

Umbach, R., Henry, N., Beard, G. F., & Berryessa, C. M. (2024). Non-consensual synthetic intimate imagery: Prevalence, attitudes, and knowledge in 10 countries. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–20. https://doi.org/10.1145/3613904.3642382

UNESCO. (2023). *Guidelines for the governance of digital platforms: Safeguarding freedom of expression and access to information through a multistakeholder approach*. https://doi.org/10.54675/OEAJ8758

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society, 6*(1), 205630512090340. https://doi.org/10.1177/2056305120903408

Ventura, G. (2024). The current state of deepfake laws in Arizona. *Arizona State Law Journal*.https://arizonastatelawjournal.org/2024/10/01/the-current-state-of-deepfake-laws-in-arizona/

Weikmann, T., & Lecheler, S. (2024). Cutting through the hype: Understanding the implications of deepfakes for the fact-checking actor-network. *Digital Journalism, 12*(10), 1505–1522. https://doi.org/10.1080/21670811.2023.2194665

Whitmarsh, L. (2011). Responsibility for justice. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195392388.001.0001

Wolff, N. E., & Safran, E. (2024, October 30). California expands its post-mortem right of publicity law to cover AI digital replicas. *CDAS*. https://cdas.com/california-expands-its-post-mortem-right-of-publicity-law-to-cover-ai-digital-replicas/

World Health Organization. (2020). *Managing the COVID-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation*. WHO Policy Brief.

Yazdani, S., Singh, A., Saxena, N., Wang, Z., Palikhe, A., Pan, D., Pal, U., Yang, J., & Zhang, W. (2025). Generative AI in depth: A survey of recent advances, model variants, and real-world applications. *Journal of Big Data, 12*, 230. https://doi.org/10.1186/s40537-025-01247-x

Young, I. M. (2011). *Responsibility for justice*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195392388.001.0001